The final publication is available at Springer, via http://dx.doi.org/10.1007/978-3-642-28885-2_3

Coordination of *-mente* ending adverbs in Portuguese: an integrated solution

Jorge Baptista[1], Lucas Nunes Vieira[2], Cláudio Diniz[3], and Nuno Mamede[4]

[1] Universidade do Algarve / Faro, Portugal
Spoken Language Lab, INESC-ID Lisboa / Lisboa, Portugal
jbaptis@ualg.pt
[2] Universidade do Algarve / Faro, Portugal
Spoken Language Lab, INESC-ID Lisboa / Lisboa, Portugal
Université de Franche-Comté / Besançon, France
lucasnvieira@gmail.com
[3] Spoken Language Lab, INESC-ID Lisboa / Lisboa, Portugal
cfpdiniz@gmail.com
[4] Instituto Superior Técnico, Universidade Técnica de Lisboa / Lisboa, Portugal
Spoken Language Lab, INESC-ID Lisboa / Lisboa, Portugal
Nuno.Mamede@inesc-id.pt

Abstract. Portuguese *-mente* ending adverbs constitute a large, morphologically homogenous, but syntactically and semantically diverse lexical set. When coordinated, the first adverb loses the adverbial suffix and takes the shape of the base adjective, in the feminine-singular form. This raises the issue of its part-of-speech (POS) classification (adverb or adjective?), but especially its adequate parsing, since it may then be incorrectly analyzed as a modifier of a preceding noun. However, the POS tagging can not be adequately performed prior to some minimal syntactic analysis. The size of the lexicon involved (more than 7,000 adverbs) and the scarcity of instances even in large corpora, make it ineffective to leave only for the POS tagger the task of solving this adjective/reduced adverbial form ambiguity. This paper proposes an integrated solution, where a rule-base disambiguating module and a POS statistical tagger combine to produce more accurate tagging and better parsing results to this non-trivial empirical problem. The system was evaluated on a large-sized corpus.

Keywords: Adverb, Coordination, POS disambiguation, Parsing, Dependency

1 Introduction

Adverbs are a significant part of the lexicon of many languages and they occur very frequently in texts. Table 1 shows the frequency of *-mente* ending adverbs (henceforward, *Adv-mente*) in two large, publicly available, corpora of Portuguese, namely the CETEMPúblico[19][5], for European Portuguese, and NILC/São Carlos [17][6], for Brazil-

[5] http://www.linguateca.pt/cetempublico/ [last access: 2012-01-12].
[6] http://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS [last access: 2012-01-12].

ian Portuguese. Even though they represent little more than 10% of all (simple) adverb occurrences in the corpora, *-mente* ending adverbs constitute the majority of the simple-word lemmas from this category [7].

	NILC/São Carlos	CETEMPúblico
lemmas (l)	397 K	1,2 M
words (w)	32,3 M	191,6 M
Adv (l)	2,867	5,361
Adv (w)	1,5 M	9,1 M
Adv-mente (l)	1,936	4,654
Adv-mente (w)	103,6 K	1,0 M

Table 1. Adverbs in two Portuguese corpora: (l) lemmas, (w) words.

When coordinated, Portuguese *-mente* ending adverbs drop the suffix and appear in the feminine-singular (*fs*) form of the base adjective:

(1a) *O Pedro leu isso lenta e atentamente* 'Peter read that slow_*fs* and attentively'
= (1b) *O Pedro leu isso lenta[mente] e [o Pedro leu isso] atentamente* 'Peter read that slow(ly)_*fs* and (Peter read that) attentively'
If there is a feminine-singular noun before the reduced adverb, it is very likely that the adverb would be considered as an adjective instead, and treated as a modifier of that noun, e.g. *a revista lenta*, 'the magazine slow' in the example below:
(2) *O Pedro leu a revista lenta e atentamente* 'Peter read the magazine_*fs* slow_*fs* and attentively'
Finally, as coordination can be iterated, longer chains of reduced adverb forms can be found:
(3) *O Pedro leu isso lenta, pausada e atentamente* 'Peter read that slow_*fs*, pausing_*fs* and attentively'
However, in both corpora, longer chains are rare. In the European Portuguese corpus mentioned above, only 24 multiple coordinated adverbs were found, against 438 simple coordination cases.

Because the reduced form of the adverb and the feminine-singular form of its base adjective are homographs, the POS of the word has to be disambiguated. However, without semantic (distributional) information on noun-adjective combinations, adverb combinations, or even verb-adverb pairs, any solution to this non-trivial problem is just an approximation.

On the other hand, it would be useless (and eventually hampering to a system) to consider that all feminine-singular adjectives could be adverbs in every context. So this particular type of strictly local ambiguity should be solved prior to general parsing rules or statistical models be applied to the text.

The performance of statistical POS taggers depends on the granularity of the tag set used by the learning algorithms, and since many systems only use a coarse tag set,

[7] Excluding compound adverbs, naturally, which are at least as numerous as simple adverbs, and also occur quite frequently in texts [10][15].

i.e., considering only the major POS category, but discarding the inflection, it is very difficult to train models sensitive to this particular phenomenon.

Finally, the coordination of adverbs, while a relatively common phenomenon in Portuguese, occurs very infrequently in texts. For the system here used, the statistical POS tagger [18], based on the Viterbi algorithm, uses a manually annotated corpus of 250K words. In this corpus only 10 instances occur of the pattern corresponding to the coordination of *Adv-mente* but only 4 are in fact coordinated *Adv-mente*. The sparsity of the phenomenon makes it an interesting challenge to NLP systems, difficult to tackle by a purely machine-learning approach. An alternative solution should be devised.

To the best of our knowledge, no assessment has been made for Portuguese on the accuracy of any disambiguation method in dealing with this specific linguistic phenomenon. The study of [1] was a preliminary survey aimed at developing the parsing rules to be implemented in the system PALAVRAS [3]. In the manually corrected and revised Portuguese treebank Bosque (version 8)[8], only 6 instances were found among 9,368 sentences. In those sentences, both the reduced form, tagged as an adjective, and the adverb are coordinated modifiers of the the same word; apparently, the only two instances of the compound *pura e simplesmente* (purely and simply) are treated as coordinated simple words; in most cases the syntactic dependency ADVL (adverbial adjunct) applies to both items, the adjective is linked to the adverb and this to the verb it modifies. While the rules themselves could not be consulted, the processing of examples (1) to (3) by the PALAVRAS parser[9] correctly yields the syntactic dependency ADVL both for (1) and (2) reduced adverb forms (lema *lento*, with the tag *mente*) modifying the verb in both (1) and (2); for (3), only the second reduced form is correctly analyzed, like in the latter examples, but *lenta* seems to be parsed as an ordinary adjective, and a PRED dependency on the root node is extracted. On the other hand, the LX-GRAM Dependency Parser[5][10], maybe due to an incorrect tagging of *lenta* as a (sg. masc.?) adjective, produces poorer results: it extracts a PRD dependency between the reduced form and the verb while the *Adv-mente* is linked through an M (modifier?) relation; naturally, the coordination between the reduced forms and the *Adv-mente* is not established.

This paper addresses the issues mentioned above in the context of the development of the STRING system [12], a Portuguese NLP chain developed at L2F/INESC ID Lisboa[11]. The system is composed of several modules, including a tokenizer, a morphological analyzer LEXMAN [7][8], a statistical POS tagger MARV [18], and a parser XIP (Xerox Incremental Parser) [2]. XIP is a cascade, finite-state, rule-based parser that analyzes sentences into chunks, extracts syntactic dependencies between chunks and it is also used for named entity recognition [11][14] and (partially) to co-reference resolution [13] and relation extraction [20].

The related problems of correct identification of the reduced adverbial form and of the parsing of coordinated *Adv-mente* is mainly a function of the morphological

[8] http://www.linguateca.pt/Floresta/corpus.html#bosque [last access: 2011-11-04].
[9] http://beta.visl.sdu.dk/visl/pt/parsing/automatic/dependency.php [last access: 2012-01-12].
[10] http://lxcenter.di.fc.ul.pt/services/en/LXServicesParserDep.html [last access: 2012-01-12].
[11] http://string.l2f.inesc-id.pt

analyzer and of the chunking module of the parser, but it is set in the more general task of extracting the syntactic dependencies between the sentences' constituents.

The paper is structured as follows: Section 2 firstly sketches the integrated solution here proposed and then presents the methods used to implement it. These involve extending lexical coverage (2.1), building new linguistically motivated rules for POS disambiguation (2.2), and constructing specific chunking (2.3) and dependency extraction (2.4) rules. Finally, to evaluate the system performance, a corpus has been built and manually annotated (2.5). Section 3 presents the evaluation of each one one of these main components of the system, while section 4 discusses these results and projects future work.

2 Methods

The strategy for the disambiguation and parsing of the coordinated *-mente* ending adverbs consists in three steps: (i) at the lexical/morphological level, instead of considering all feminine adjectival forms as adverbs, extend the coverage of the existing lexicon of adverbs, and associate them to their reduced forms; (ii) use this morphologic information with that of the environing words in order to build linguistically motivated rules and locally determine the patterns where a coordination of adverbs is likely to occur or, on the contrary, where a reduced adverbial form can reasonably be discarded; at the end of this rule-based, disambiguation process, all remaining ambiguous forms are tagged as adjectives; (iii) based on the results from previous steps, produce the adequate chunking and extract the syntactic-semantic dependencies between the sentence constituents.

In the next subsections, these processing steps are described in detail. For the evaluation, a corpus with 1,132 sentences was collected from the CETEMPúblico, containing instances of coordination of a (surface) feminine-singular adjective or past participle with an *Adv-mente*. The corpus was parsed by the system and the output was manually corrected by two linguists. In the last subsection, corpus collection and annotation will be briefly presented.

2.1 Lexicon

The existing lexicon of the system has been systematically completed by adding all *Adv-mente* entries found in an orthographic vocabulary [4]. These correspond to 3,614 entries. Then, all valid *-mente* ending forms found in the European Portuguese corpus were manually perused and the adverbs selected. Duplicates from the first list were removed, thus yielding 3,636 new entries. For each entry, the feminine-singular form of the base adjective was automatically generated and the list was then manually revised for errors and for the insertion of orthographic variants, resulting from the new, unified Portuguese orthography. The final list consists of 7,250 *-mente* ending adverbs. For example, the entry for *abstratamente* 'abstractly' is associated with the orthographic variant *abstractamente*, and to the reduced forms *abstrata* and *abstracta* 'abstract_*fs*'. This reduced form is then given the feature '`r`' (from 'reduced'). When analyzing a sentence where *abstracta* appears, at this morphologic stage, the system produces the

following tags (format adapted for clarity):

```
abstracta: abstratamente Adv_r; abstrata Adj_fs
```

In this way, only forms with attested *-mente* adverbial counterparts are validated.

It has been previously noted by [1] that compound adverbs (or colocational combinations), such as *única e exclusivamente* 'uniquely and exclusively' and *única e simplesmente* 'uniquely and simply' occurred quite often in the corpus. To these forms, others were added in the lexicon, v.g. *pura e simplesmente* 'purely and simply', *dire(c)ta ou indire(c)tamente* 'directly or indirectly', *explícita ou implicitamente* 'implicitly or explicitly' and *total ou parcialmente* 'totally or partially'. These combinations occur 3,074 times in the CETEMPúblico. In our corpus, only *pura e simplesmente* occurs, 220 times.

2.2 Rule-based disambiguation

The next step in the system processing chain is a rule-based disambiguation module [7],[8]. The linguistically motivated disambiguation rules produced are at the core of the solution here presented. These rules are regular expressions that take the general form:

```
<left-context>|<pattern>|<right-context> := <result>
```

where <pattern> corresponds to the ambiguous target word and the different categories it may be associated with; <result> consists in selecting (+) or discarding (−) a given category; the left and right contexts are facultative. For example, the general rule below selects the adverb reduced form when it appears coordinated with a *-mente* ending adverb:

```
0> [CAT='adv',SYN='red'][CAT='adj'] |
[surface='e'];[surface='ou'];[surface='mas'],
[surfaceRegex='.+mente',CAT='adv'] |
:= [CAT='adv']+.
```

This rule reads as follows: the left context is empty; the <pattern> consists of the ambiguous form adverb/adjective; the adverbial form must present the feature SYN with the value 'red' (for 'reduced'); then follows the right context, where the coordinative conjunctions and the *Adv-mente* are explicit; for the conjunctions, the surface form is sufficient; to define the adverb, a regular expression is used along with its POS.

Most rules have to be duplicated in order to deal with the feminine-singular form of past participles. This is the purpose of the rule below:

```
0> [CAT='adv',SYN='red'][MOD='par',GEN='f',NUM='s'] |
[surface='e'];[surface='ou'];[surface='mas'],
[surfaceRegex='.+mente',CAT='adv'] |
:= [CAT='adv']+.
```

Rule-order application is fixed, so more specific rules are stated before more general ones. For example, the pattern of coordinated adjectives, each modified by an adverb, is more constrained than the previous patterns and it is thus stated before the general rules above:

```
0> [CAT='adv'] |
[CAT='adv',SYN='red'][CAT='adj',GEN='f',NUM='s'] |
[CAT='con',SCT='coo'],[surfaceRegex='.+mente',CAT='adv'],
```

```
[CAT='adj',GEN='f',NUM='s'][MOD='par',GEN='f',NUM='s'] |
:= [CAT='adv']-.
```

Some rules require lists of words to be spelled out, such as the next one, where a negation adverb in front of an ambiguous adjective is the context that allows to discard the reduced adverbial form; the negation adverb is provided by a list of words (at later stages, namely in the parser, this piece of information is encoded by way of feature-value pairs):

```
0> |[surface='não'];[surface='nem'];[surface='nunca';
[surface='jamais'];[surface='nada'] |
[CAT='adv',SYN='red'][CAT='adj'] |
[surface='e'];[surface='ou'];[surface='mas'],
[surfaceRegex='.+mente',CAT='adv'] |
:= [CAT='adv']-.
```

Finally, at the last stage of the process and for the remaining ambiguous forms, the tag corresponding to the reduced adverb form is discarded by a general "cleaning" rule:

```
0> [CAT='adv',SYN='red'][SYN= 'red']
:= [SYN='red']-.
```

So far, 16 rules have been devised, based on known cases of ambiguity. Our approach is conservative, in the sense that rules tend to be general in scope and as much precise as possible. During this process, some rules were devised but not yet implemented, for they are not linguistically well motivated even though the patterns appear often in text. This is the case of coordinated adjectives after a copula verb, where the second adjective is modified by an *Adv-mente*, as in (4):

(4) *A crítica portuguesa foi agressiva e extremamente injusta*

'The Portuguese critic_*fs* (=the critics) was aggressive and extremely unfair'

or when both adjectives are modified, especially if a quantifying adverb is involved, as in (5):

(5) *A corrida também é muito longa e fisicamente dura*

'The race also is very long and physically hard'

Strictly speaking, these patterns are grammatically ambiguous, but more often than not the adjective is found in this context.

2.3 Chunking

In the chunking stage, the XIP parser analyzes the sentence by splitting it into elementary constituents (or chunks). Ordinarily, a stand-alone adverb construes an adverbial phrase (`ADVP`). Chunks are formed according to chunking rules, such as the following, allowing up to three consecutive adverbs to form an `ADVP`:

```
ADVP @= (adv), (adv), adv.
```

At this stage, the system can make use of a rich set of lexicons, featuring syntactic and semantic information, as well as the information derived from the morphological analyzer. In the coordination of *Adv-mente*, an `ADVP` is construed. For example, for sentence (1) the following chunking is produced:

```
0>TOP{NP{O Pedro} VF{leu} NP{isso}
ADVP{lenta e atentamente} .}
```

This ADVP results from the application of the following rule:

```
18> ADVP @=   |?[noun,fem,sg] |
(adv[advquant];adv[advcomp];adv[neg])*,
adv[reducedmorph],
conj[lemma:e];conj[lemma:ou];conj[lemma:mas],
(adv[advquant];adv[advcomp];adv[neg])*,
adv[surface:"%c+mente"].
```

The chunking rule reads: an ADVP chunk is built with two coordinated adverbs, the first is a reduced form, indicated by the feature [reducemorph], and the second an *Adv-mente*; only the conjunctions *e* (and), *ou* (or) and *mas* (but) are allowed; both adverbs can facultatively be further modified by a quantifying adverb, a comparative adverb or a negation adverb; these adverbs have been given in the lexicon the features [advquant], [advcomp] and [neg], respectively; this chunking is not made if there is a feminine-singular noun in the left context of the pattern. A similar rule is used for coordination of three (or more) *Adv-mente*.

2.4 Dependency Extraction

Finally, the parser extracts the syntactic relations between the chunks. Dependency extraction rules have the general format:

```
 <left-context> |<pattern> |<right-context>
if <conditions> <dependencies>
```

Relevant for this paper are the *coordination* (COORD) and *modifier* (MOD) dependencies, which are now very briefly presented.

Coordination is a strictly local relation between a coordinative conjunction and two (or more) chunk heads. This dependency is extracted as early as possible in the parsing process, and before the modifier is calculated. In the case of the coordination of *Adv-mente*, the COORD dependency between the reduced form is given the feature c-mente. The basic rule for coordination extraction is provided below:

```
|ADVP{?*, adv#1,
conj#2[lemma:e];conj#2[lemma:ou];conj#2[lemma:mas],
(adv[advquant];adv[advcomp];adv[neg])*, adv#3[last]} |
if ( CLINK(#1,#3))
CLINK(#1,#3),
LCOORD[c-mente=+](#2,#1),
RCOORD(#2,#3).
```

The rule reads: in an ADVP chunk with two coordinated adverbs, which are designated by variables #1 and #3, if no auxiliary dependency CLINK has yet been extracted between the two adverbs; then create that CLINK dependency between the adverbs, that is, the conjunction proper; and create two other dependencies, to produce an output: LCOORD (L=left) between the conjunction (variable #2) and the reduced form, which is then given the feature c-mente, and RCOORD (R=right) between the conjuncion and the *Adv-mente*. A facultative (quantifier, comparative or negation) adverb can occur between the conjunction and the *Adv-mente*; the modifier relation holding between these two adverbs is extracted by another rule.

For longer coordination chains, involving two or more reduced forms, the LCOORD dependency is propagated to the left by a similar rule:

```
|ADVP{?*, adv#1,
conj#2[lemma:e];conj#2[lemma:ou];conj#2[lemma:mas],
if (~CLINK(#1,#3) & CLINK(#3,?) & LCOORD(#2,#3))
CLINK(#1,#3),
LCOORD[c-mente=+](#2,#1).
```

The modifier dependency holds between two chunks. For *Adv-mente*, most of them modify a verb or an adjective. One of the basic rules for extracting the adverbial, right modifier of a verb is given below:

```
|#1[verb];sc#1, ?[verb: ~ ,scfeat: ~ ],
(AP;PP), (PUNCT[comma]), ADVP#2 |
if ( HEAD(#3,#1) & HEAD(#4,#2) & ~ MOD(?,#4)
& ~ QUANTD(#3,#4))
MOD[post=+](#3,#4)
```

Briefly, this rule reads: For a verb (or a subclause SC) #1 and an adverbial phrase #2, eventually admitting an adjectival or prepositional phrase, or a comma, in between; if no modifier MOD has been extracted for the head of #2, nor a quantifier QUANTD dependency has been extracted between the heads of #1 and #2; then build the MOD dependency between the heads of the verb and the adverb phrases.

The result of the dependency extraction process for sentence (1) *O Pedro leu isso lenta e atentamente* 'Peter read this slowly and attentively' is the following:

```
MAIN(leu)                MOD_POST(leu,atentamente)
DETD(Pedro,O)            MOD_C-MENTE_POST(leu,lenta)
COORD_C-MENTE(e,lenta)   SUBJ_PRE(leu,Pedro)
COORD(e,atentamente)     CDIR_POST(leu,isso)
VDOMAIN(leu,leu)         NE_PEOPLE_INDIVIDUAL(Pedro)
```

Briefly, the dependencies above include the subject (SUBJ) and direct object (CDIR); the determinant (DETD) and the named entity (NE); the main (MAIN) element of the sentence; the verb domain (VDOMAIN), for dealing with auxiliary verbal chains (not relevant in this example); and, finally, the two coordination dependencies involving the adverbs, and the corresponding modifier dependencies. Features _PRE and _POST indicate if the dependent is to the left or to the right of the dependency head.

2.5 The evaluation corpus

For the evaluation, a corpus, with 1,132 sentences, was retrieved from the CETEMPúblico. It consists of sentences presenting an adjective or past participle, one of the three main coordinating conjunctions – *e* (and), *ou* (or) or *mas* (but) –, and an *Adv-mente*. The sentences were obtained from the concordances retrieved using the AC/DC search system of Linguateca webpage. The corpus was then parsed by the system and the dependencies

were manually corrected, each sentence being independently checked at least twice, by two linguists. The chunking was also corrected, when appropriate. For this paper, only the COORD and MOD dependencies involving *Adv-mente* or their reduced forms were kept from the system's output. Table 2 shows the breakdown of each dependency in the corpus. The difference between COORD and COORD_C-MENTE is due to the cases of multiple coordination (i.e., more than two adverbs coordinated together). The large difference between MOD and MOD_C-MENTE consist of *Adv-mente* that, although occurring next to a conjunction and after a reduced form, are not coordinated with it, and modify some other constituent in the sentence.

Table 2. Dependencies in Reference Corpus

Dependency	#
COORD	438
COORD_C-MENTE	462
MOD	1,403
MOD_C-MENTE	462

3 Evaluation and Results

To assess the integrated solution implemented in the system, each of its three main steps was evaluated independently. A set of scripts were especially built to make the result-gathering process fully automatic.

3.1 Lexicon

First, the lexicon coverage is evaluated by computing the recall of the reduced forms. From the 462 reduced forms found in the reference corpus, only 10 (8 different) forms had not been previously encoded in the lexicon, thus yielding a recall of 0.978, *scilicet*: Two so-called point-of-view adverbs [16] (*bioquímica* 'biochemistry' and *iconográfica* 'iconographic'), four participle-based (*figurada*, *fundada*, *interpelada*, *zelada*), one numeral based (dupla 'double'), and a spelling mistake (*massiça* = *maciça* 'massive'). The numeral-based form is clearly a lacuna, since not only the *Adv-mente* is already in the lexicon (*duplamente* 'doubly', but other reduced forms also have been encoded (*tripla*, *triplamente* 'triple, three times'). The remaining lacunae were also corrected. For the misspelled form and its variants, because this is a very frequent error in texts, a new entry, but with the correct lemma, was introduced in the lexicon[12].

[12] A finite-state morphological analyzer is currently under construction, to complement LexMan [7][8].

3.2 Disambiguation Rules

This step consists in assessing the impact of the disambiguation rules in selecting or discarding the POS tags corresponding to the adjective or the reduced adverbial form. Table 3 shows the results of the rule-based disambiguation module. From the 462 adverb reduced forms, the system fails to spot 21, while it incorrectly accords this tag to 316, therefore yielding a relatively low precision but high recall, contributing to the interesting F-measure result. This means that in spite of the conservative approach in devising the disambiguation rules and the final, "cleaning" rule that eliminates all remaining reduced forms not previously captured, the system still fails to recognize the cases where there is no coordination of adverbs.

Table 3. Results: Disambiguation Rules

Precision	Recall	F-Measure
0.583	0.955	0.724

3.3 Dependency Extraction

The next figures are a combined result of the chunking and of the dependency extraction modules. The purpose of parsing a text is to retrieve the syntactic-semantic relations between constituents, which (partially) express the text meanings. Table 4 shows the results for the dependency extraction module. In order to obtain a better perception of the system performance, a set of experiments was carried out. The first line presents the overall performance of the system. In the next lines, each dependency is evaluated separately. Finally, the two coordination and modifier dependencies are evaluated in pairs.

Table 4. Results: Dependency Extraction

Experiment	Precision	Recall	F-Measure
All dependencies	0.754	*0.875*	0.810
MOD	*0.921*	0.852	*0.886*
MOD_C-MENTE	0.608	0.719	0.659
COORD	0.642	0.777	0.703
COORD_C-MENTE	0.646	0.805	0.717
2MOD	0.822	0.849	0.834
2COORD	0.644	0.858	0.736

The overall performance of the system in the dependency extraction is promising. In general, the system is able to extract most of the modifier dependencies (92%), and

only 39% of reduced adverbial forms are not adequately related to the element they modify. The system shows suboptimal performance in the extraction of coordination dependencies. There is a clear relation between the low precision in the MOD_C-MENTE and the low precision on COORD dependencies. When the system fails to extract the coordination, it also (partially) fails to extract the modifiers. The reason for this is to be found in the previous module of disambiguation rules, which often and inadequately selects the reduced adverb form instead of recognizing the coordination of adjectives.

4 Discussion and Future Work

These results confirm the difficulty of the task, sketched at the onset of this paper. The overall performance of the system may be considered satisfactory. The linguistic resources of the STRING natural language processing chain were systematically extended to be used in this paper and they proved to be comprehensive showing very good lexical coverage, and featuring a 0.98 recall.

Adequately capturing coordination is a difficult parsing task, mainly because of the different sentential levels at which it may operate, but also because of the many semantic constraints involved in the pairing of two constituents. In the case of reduced adverbs, ambiguity with another part-of-speech complicates matters even further, lowering results.

Precision in the dependency extraction, while generally good (0.75), is directly related to the low precision of disambiguation rules (0.58), which needs to be improved. The main cause for this low precision is the excessive tendency to analyze adjectives as adverbs in coordination. This could be avoided by using disambiguating rules which would be linguistically less motivated, but that would be more in accordance with the patterns frequently found in the corpus. For example, with coordinated adjectives, the presence of a copula verb often occurs (e.g. *a sua confusão é normal e provavelmente resolve-se com a experiência* 'his confusion is normal and probably can be solved with the experience'); the same happens in the presence of a quantifying adverb before the first adjective (e.g. *uma região bastante conservadora e notoriamente católica*), or the second, or both adjectives; there is also a tendency for the last adjective in a sequence of three coordinated adjectives to present an adverb modifier (e.g. *uma coisa horrível, ilegal e altamente reprovável* 'something horrible, illegal and highly reproachable'). Such solution, however, risks not to be easily adaptable to other domains or text genres. Another path to be tread would consist of using available semantic and syntactic information associated to *Adv-mente* [9], and collocational patterns they may show [21] to model the correct classification, using machine-learning techniques.

References

1. Afonso, S.: *Clara e sucintamente*: um estudo em corpus sobre a coordenação de advérbios em *-mente*. In: *XVIII Encontro Nacional da Associação Portuguesa de Linguística* (APL 2002). pp. 27–36. Porto, Portugal (2002)

2. Aït-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness beyond shallowness: Incremental deep parsing. *Natural Language Engineering* 8, 121–144 (2002)
3. Bick, E.: *The Parsing System PALAVRAS. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Arhus University Press (2000).
4. Casteleiro, J.M.: *Vocabulário Ortográfico da Língua Portuguesa*. Porto: Porto Editora (2009)
5. Costa, F.; Branco, A.: LX-Gram: A Deep Linguistic Processing Grammar for Portuguese. In Pardo, T. et al. (eds.), *Computational Processing of the Portuguese Language*. *Lecture Notes in Artificial Intelligence*, 6001, pp.86-89, Berlin: Springer (2010)
6. Costa, J.: *O Advérbio em Português Europeu*. Lisboa: Edições Colibri (2008)
7. Diniz, C.: RUDRICO2 – *Um Conversor Baseado em Regras de Transformação Declarativas*, M.Sc. Thesis, Lisboa: Instituto Superior Técnico/Universidade Técnica de Lisboa (2010)
8. Diniz, C.; Mamede, N.: LEXMAN - *Lexical Morphological Analyser*, Tech. rep., Lisboa L2F/INESC-ID Lisboa (2011)
9. Fernandes, G.: *Automatic Disambiguation of* -mente *ending Adverbs in Brazilian Portuguese*. M.A. Thesis, Universidade do Algarve/Universitat Autònoma de Barcelona, Faro/Barcelona (2011)
10. Gross, M.: *Grammaire transformationnelle du français. 3 - Syntaxe de l'adverbe*. ASSTRIL, Paris (1986)
11. Hagège, C.; Baptista, J.; Mamede, N.: Caracterização e Processamento de Expressões Temporais em Português. *Linguamática* 2(1): 63–76 (2010)
12. Mamede, N.: STRING – *A Cadeia de Processamento de Língua Natural do* L2F. Tech. Rep., Lisboa: L2F/INESC-ID Lisboa (2011)
13. Nobre, N.: *Anaphora Resolution*. M.Sc. Thesis, Lisboa: Instituto Superior Técnico/Universidade Técnica de Lisboa (2011)
14. Oliveira, D.: *Extraction and Classification of Named Entities*, M.Sc. Thesis, Lisboa: Instituto Superior Técnico/Universidade Técnica de Lisboa (2010)
15. Palma, C.: *Expressões fixas adverbiais: descrição léxico-sintáctica e subsídios para um estudo contrastivo Português-Espanhol* (M.A. Thesis). Faro, Univ. Algarve - FCHS, Faro, Universidade do Algarve/FCHS (2009)
16. Molinier, C. and Levrier, F.: *Grammaire des Adverbes. Description des formes en* -ment. Librairie Droz, Genève-Paris (2000)
17. Pinheiro, G.M., Aluísio, S.M.: *Corpus NILC: Descrição e análise crítica com vistas ao projeto Lácio-Web*. Tech. rep., NILC-TR-03-03, São Carlos, Brasil (Fevereiro 2003)
18. Ribeiro, R.: *Anotação Morfossintáctica Desambiguada em Português*, MSc Thesis, Lisboa: Instituto Superior Técnico/Universidade Técnica de Lisboa (2003)
19. Rocha, P., Santos, D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada*, PROPOR'2000. pp. 131–140. Atibaia, São Paulo, Brasil (November 2000)
20. Santos, D.: *Extracção de Relações entre Entidades Mencionadas*, MSc Thesis, Lisboa: Instituto Superior Técnico/Universidade Técnica de Lisboa (2010)
21. Vieira, L.: *Verb and* -mente *ending Adverb Collocations in Brazilian Portuguese: Extraction from Corpora and Automatic Translation into English*. M.A. Thesis (in preparation).